Teste de Kolmogorov-Smirnov (Uma amostra)

Professor: Pedro M. Almeida-Junior

23 de julho de 2021

Departamento de Estatística (UEPB)

O teste de Kolmogorov-Smirnov

- O teste de Kolmogorov-Smirnov de **uma amostra** é outro **teste de aderência**.
- Mede o grau de concordância entre a distribuição de um conjunto de valores da amostra (escores observados) e alguma distribuição teórica especificada.
- O teste envolve especificar a distribuição de frequência acumulada que ocorreria dada a distribuição teórica e compará-la com a distribuição de freqüência acumulada observada.
- A distribuição teórica representa o que seria esperado sob H_0. O ponto no qual essas duas distribuições, teórica e observada, mostram a maior divergência é determinado.

- A distribuição amostral indica a possibilidade de que ocorresse uma **divergência** da magnitude observada se as observações fossem realmente uma amostra aleatória de uma distribuição teórica.
- O **teste de Kolmogorov-Smirnov** admite que a distribuição da variável que está sendo testada é contínua, como especificado pela distribuição de frequências acumuladas.
- O teste é apropriado para **testar a aderência** para variáveis que são medidas pelo menos em uma escala ordinal.
- **Pressuposto:** As variáveis que são medidas tem que estar pelo menos em uma escala ordinal

Método

- Seja $F_0(X)$ uma função completamente especificada de distribuição de frequências relativas acumuladas (a distribuição teórica sob H_0).

- Para qualquer valor de X, o valor de $F_0(X)$ é a proporção de casos esperados com escores iguais ou menores do que X.

- $S_N(X)$ é a distribuição de frequências relativas acumuladas observada de uma amostra aleatória de N observações.

- Se X_i é um escore qualquer possível, então $S_N(X_i) = Q_i/N$, onde Q_i é o número de observações menores ou iguais a X_i .

- $F_0(X_i)$ é a proporção esperada de observações menores ou iguais a X_i.

- Quando H_0 é verdadeira, esperaríamos que as diferenças entre $S_N(X_i)$ e $F_0(X_i)$ fossem pequenas e dentro dos limites de erros aleatórios. O teste de Kolmogorov-Smirnov focaliza sobre o maior dos desvios. O maior valor absoluto de $F_0(X_i) - S_N(X_i)$ é chamado de desvio máximo D :

$$D = \max |F_0(X_i) - S_N(X_i)| \quad i = 1, 2, \ldots, N$$

Distribuição amostral

- A distribuição amostral de D sob H_0 é conhecida. A Tabela exibida a seguir fornece certos valores críticos da distribuição amostral.

- Suponha que um pesquisador tenha uma amostra de tamanho $N = 43$ e escolha $\alpha = 0,05$. A Tabela mostra que qualquer $D \geq 1,36/\sqrt{N}$ será significante. Isto é, qualquer D, que seja maior ou igual a $1,36/\sqrt{43} = 0,207$ será significante no nível 0,05 (teste bilateral).

Tamanho da amostra (N)	Nível de significância para $D = \text{máximo } \lvert F_0(X) - S_N(X) \rvert$				
	0,20	0,15	0,10	0,05	0,01
1	0,900	0,925	0,950	0,975	0,995
2	0,684	0,726	0,776	0,842	0,929
3	0,565	0,597	0,642	0,708	0,828
4	0,494	0,525	0,564	0,624	0,733
5	0,446	0,474	0,510	0,565	0,669
6	0,410	0,436	0,470	0,521	0,618
7	0,381	0,405	0,438	0,486	0,577
8	0,358	0,381	0,411	0,457	0,543
9	0,339	0,360	0,388	0,432	0,514
10	0,322	0,342	0,368	0,410	0,490
11	0,307	0,326	0,352	0,391	0,468
12	0,295	0,313	0,338	0,375	0,450
13	0,284	0,302	0,325	0,361	0,433
14	0,274	0,292	0,314	0,349	0,418
15	0,266	0,283	0,304	0,338	0,404
16	0,258	0,274	0,295	0,328	0,392
17	0,250	0,266	0,286	0,318	0,381
18	0,244	0,259	0,278	0,309	0,371
19	0,237	0,252	0,272	0,301	0,363
20	0,231	0,246	0,264	0,294	0,356
25	0,21	0,22	0,24	0,27	0,32
30	0,19	0,20	0,22	0,24	0,29
35	0,18	0,19	0,21	0,23	0,27
Acima de 50	$\frac{1{,}07}{\sqrt{N}}$	$\frac{1{,}14}{\sqrt{N}}$	$\frac{1{,}22}{\sqrt{N}}$	$\frac{1{,}36}{\sqrt{N}}$	$\frac{1{,}63}{\sqrt{N}}$

Figura 1: Valores Críticos de D no teste de uma amostra Kolmogorov-Smirnov.

Passos para o Teste de Kolmogorov-Smirnov

Na **aplicação do teste de Kolmogorov-Smirnov**, os passos são os seguintes:

1. Especifique a distribuição acumulada teórica, isto é, a distribuição acumulada esperada sob H_0
2. Organize os escores observados em uma distribuição acumulada e converta as frequências acumuladas em frequências relativas acumuladas $[S_N(X_i)]$. Para cada intervalo encontre a frequência relativa acumulada esperada $F_0(X_i)$.
3. Encontre D.
4. Consulte a tabela com os valores críticos de D para encontrar a probabilidade (bilateral) associada com a ocorrência sob H_0 de valores tão grandes quanto o valor observado de D.
5. Se essa probabilidade é igual ou menor que α, rejeite H_0.

Exemplo 1: Durante vários anos os pesquisadores têm estudado a duração de uma variedade de eventos tais como empregos, greves e guerras. Como uma parte de tal pesquisa, suposições precisas concernentes a ações individuais e ao curso dos eventos têm levado a modelos matemáticos para os eventos, os quais fazem predições sobre suas distribuições. Enquanto detalhes dos modelos matemáticos não são de interesse especial aqui, o conhecimento sobre a concordância entre os dados e as predições do modelo proporciona uma boa ilustração do teste de Kolmogorov-Smirnov de aderência de uma amostra. Os dados concernentes à duração das greves que começaram em 1965 no Reino Unido foram coletados, analisados, e predições foram feitas com o uso do modelo matemático. A Tabela seguinte contém a distribuição de frequências acumuladas das durações de $N = 840$ greves. Também estão dadas na tabela as frequências acumuladas preditas pelo modelo matemático.

Exemplo 1: formulação do problema

1. **Hipótese nula:** H_0: A diferença entre as durações observada e predita não excede as diferenças que seriam esperadas que ocorressem devido ao acaso e H_1 : as durações observadas das greves não coincidem com as preditas pelo modelo matemático.

2. **Teste estatístico:** O teste de Kolmogorov-Smirnov de uma amostra é escolhido porque o pesquisador deseja comparar uma distribuição observada de escores de uma escala ordinal com a distribuição teórica de escores.

3. **Nível de significância:** Seja $\alpha = 0,05$ e N o número de greves que começaram no Reino Unido em 1965, então, $N = 840$.

4. **Distribuição amostral:** O valores críticos de D e o desvio absoluto máximo entre as distribuições acumuladas observada e predita são tabelados, junto com suas probabilidades associadas de ocorrência quando H_0 é verdadeira.

5. **Região de rejeição:** A região de rejeição consiste de todos os valores de D , que sejam tão grandes que a probabilidade associada com suas ocorrências quando H_0 é verdadeira seja menor ou igual a $\alpha = 0,05$.

5. **Decisão:** O valor de D, a diferença máxima entre as frequências acumuladas é $| F_0(X) - S_N(X)| = |510,45/840 - 523/840 |= 0,015$. Como $N > 35$, devemos usar aproximações para amostras grandes. Com $N = 840$ o valor crítico de D é $1,36/\sqrt{840} = 0,047$. **Qual decisão final ?**

Duração máxima (dias)	Freqüência acumulada		Freqüência relativa acumulada		$\lvert F_0(X) - S_N(X) \rvert$
	Observada	Predita	Observada	Predita	
1 – 2	203	212,81	0,242	0,253	0,011
2 – 3	352	348,26	0,419	0,415	0,004
3 – 4	452	442,06	0,538	0,526	0,012
4 – 5	523	510,45	0,623	0,608	0,015
5 – 6	572	562,15	0,681	0,669	0,012
6 – 7	605	602,34	0,720	0,717	0,003
7 – 8	634	634,27	0,755	0,755	0,000
8 – 9	660	660,10	0,786	0,786	0,000
9 – 10	683	681,32	0,813	0,811	0,002
10 – 11	697	698,97	0,830	0,832	0,002
11 – 12	709	713,82	0,844	0,850	0,006
12 – 13	718	726,44	0,855	0,865	0,010
13 – 14	729	737,26	0,868	0,878	0,010
14 – 15	744	746,61	0,886	0,889	0,003
15 – 16	750	754,74	0,893	0,899	0,003
16 – 17	757	761,86	0,901	0,907	0,006
17 – 18	763	768,13	0,908	0,914	0,006
18 – 19	767	773,68	0,913	0,921	0,008
19 – 20	771	778,62	0,918	0,927	0,009
20 – 25	788	796,68	0,938	0,948	0,010
25 – 30	804	807,86	0,957	0,962	0,005
30 – 35	812	815,25	0,967	0,971	0,004
35 – 40	820	820,39	0,976	0,977	0,001
40 – 50	832	826,86	0,990	0,984	0,006
> 50	840	840,01	1,000	1,000	0,000

Figura 2: Dados Exemplo 1.

Exemplo: Verificar, ao nível $\alpha = 0,05$, se os dados abaixo se distribuem segundo uma distribuição de Poisson com parâmetro $\lambda = 3$.

X_i	0	1	2	3	4	5	6	7	Total
Freq. Observada	7	14	18	23	22	9	3	4	100

Poder do teste

- Quando **amostras são pequenas** e **categorias adjacentes precisam ser combinadas** para o uso apropriado da estatística X^2, **o teste qui-quadrado é definitivamente menos poderoso do que o teste de Kolmogorov-Smirnov**.

- Para amostras muito pequenas, o teste qui-quadrado não pode ser usado, mas o teste de Kolmogorov-Smirnov pode.

- No entanto, **é possível que os testes forneçam resultados similares**, particularmente quando o tamanho da amostra é grande.

- O teste Qui-quadrado assume que as distribuições são nominais, enquanto o teste de Kolmogorov-Smirnov assume uma **distribuição contínua**. Em princípio, ambos os testes poderiam ser aplicados a dados ordinais; entretanto, **o agrupamento necessário para a aplicação do teste qui-quadrado torna-o menos preciso do que o teste de Kolmogorov-Smirnov**.

Kolmogorov-Smirnov versus Qui-Quadrado

Com amostras pequenas, o **teste de Kolmogorov-Smirnov é exato**, enquanto o teste Qui-Quadrado de aderência é somente aproximado. Neste caso, a preferência deve ser dada ao teste de Kolmogorov-Smirnov.